Refining Reasoning Chains through Self Correcting Reinforcement Learning Architectures for Mitigating Logical Hallucinations in Large Language Models

Ethan Thornton
Department of Systems Engineering, Colorado School of Mines
e.thornton@mines.edu

Abstract

The rapid proliferation of large language models (LLMs) across critical socio-technical infrastructures has necessitated a paradigm shift from mere generative fluency to rigorous logical reliability. Despite advancements in scale, LLMs remain susceptible to logical hallucinations—instances where a model produces structurally coherent but substantively invalid reasoning chains. These failures present significant risks in domains such as legal adjudication, medical diagnostics, and engineering design, where the internal consistency of an argument is as vital as the final output. This paper proposes a systems-level architectural framework for refining reasoning chains through self-correcting reinforcement learning (RL). By integrating modular refiner policies with adaptive solver hierarchies, we transition the alignment burden from static fine-tuning to dynamic, inference-time optimization. We analyze the structural trade-offs between computational overhead and logical robustness, emphasizing the role of verifiable reward signals in stabilizing the iterative refinement process. Our discussion extends to the governance implications of deploying these architectures in public-facing systems, addressing the socio-technical challenges of transparency, fairness, and the prevention of reward hacking. Through a multi-dimensional analysis of infrastructure and policy, we argue that the future of resilient AI lies in the convergence of generative potential and autonomous corrective feedback loops, ensuring that reasoning remains grounded in verifiable logic rather than stochastic approximation.

Keywords:
Large Language Models, Logical Hallucinations, Reinforcement Learning, Self-Correction, Reasoning Chains, Socio-Technical Infrastructure, AI Governance.

1. Introduction

The evolution of generative artificial intelligence has reached a critical juncture where the primary bottleneck is no longer the breadth of information retrieval but the structural integrity of internal reasoning [23]. While contemporary large language models exhibit remarkable capabilities in natural language understanding and synthesis, their deployment in high-stakes environments is frequently hampered by the phenomenon of logical hallucinations [26]. Unlike factual hallucinations, which involve the generation of incorrect empirical data, logical hallucinations

represent a breakdown in the inferential process itself [11]. A model may correctly identify premises but fail to apply the transitive or deductive rules required to reach a valid conclusion, resulting in a reasoning chain that is superficially persuasive but fundamentally flawed [20]. This challenge is compounded by the "black box" nature of massive neural architectures, where the transition from token prediction to logical deduction remains poorly understood from a systems perspective [18].

As LLMs are increasingly integrated into the backbone of modern society—from automated code generation for critical infrastructure to policy drafting for governmental bodies—the cost of these logical failures escalates [19]. A single erroneous reasoning step in a structural engineering analysis or a legal brief can lead to catastrophic downstream consequences [21]. To address this, the research community has shifted focus toward "Chain-of-Thought" (CoT) prompting and iterative refinement [23]. However, static prompting techniques often fail to generalize across diverse task distributions and can even exacerbate confirmation bias within the model [6]. The core problem lies in the static nature of the alignment: models are often trained to mimic human-like reasoning patterns rather than to adhere to formal logical constraints [16].

This research paper explores the design and implementation of self-correcting reinforcement learning architectures specifically tailored to mitigate these logical inconsistencies. We argue that the mitigation of hallucinations requires a shift from monolithic model updates toward modular, agentic systems that can autonomously evaluate and refine their own reasoning paths [4]. By treating the reasoning process as a sequential decision-making task, we can apply reinforcement learning techniques to reward consistency, structural validity, and corrective behavior [25]. This systems-level approach considers not only the algorithmic implementation but also the broader implications for deployment, infrastructure sustainability, and the ethical governance of autonomous reasoning systems [27].

2. The Architecture of Logical Failure: Mapping Hallucinations in LLMs

Logical hallucinations are not merely errors in output; they are structural failures in the latent space of the model [2]. To understand how to refine these chains, one must first categorize the types of logical breakdowns that occur during inference. Traditional taxonomies of LLM failure often conflate factual inaccuracies with logical inconsistencies [1]. However, in a systems engineering context, these are distinct failure modes requiring different mitigation strategies. Fact-conflicting hallucinations emerge from the training data or retrieval mechanisms, whereas context-conflicting and logic-conflicting hallucinations emerge from the transformer's inability to maintain a coherent state over long-range dependencies [26].

In complex reasoning tasks, such as those found in mathematics or multi-step logic puzzles, a model may undergo what is termed "reasoning-answer misalignment" [5]. In this scenario, the model produces a chain of thought that correctly identifies the necessary steps but ultimately arrives at a wrong answer, or conversely, produces a correct answer based on flawed premises [10]. This decoupling suggests that the model is relying on spurious correlations—such as the frequency of certain tokens in the training set—rather than a robust internal representation of logic

[14]. From an architectural standpoint, this is a failure of the attention mechanism to prioritize relational logic over probabilistic token proximity.

Furthermore, the imperceptibility of these errors poses a significant challenge for human-in-the-loop systems [16]. Because LLMs are trained to be highly persuasive and fluent, their logical failures are often buried under layers of eloquent prose. This effect can lead human operators to over-rely on the system, a phenomenon known as automation bias [26]. In socio-technical infrastructures, where LLMs act as decision-support tools, the invisibility of logical hallucinations degrades the overall reliability of the human-machine team [19]. Addressing this requires a system that can not only generate reasoning but also provide an internal "audit trail" that is subject to autonomous verification [9].

3. Self-Correcting Reinforcement Learning: A Systems Perspective

The transition from static generation to self-correcting reasoning involves the implementation of a feedback-driven architecture. At the heart of this proposal is the use of Reinforcement Learning with Verifiable Rewards (RLVR) [24]. Unlike traditional Reinforcement Learning from Human Feedback (RLHF), which relies on subjective human preferences, RLVR utilizes objective markers of correctness—such as code execution results, mathematical proofs, or formal logic checks—to provide precise supervision [28]. This shift reduces the noise inherent in human labeling and allows the system to scale its learning through autonomous interaction with a structured environment [15].

A robust self-correcting architecture typically consists of a "Refiner" and a "Solver." The Refiner is a specialized meta-policy trained via RL to rewrite and decompose raw human queries into explicit logical steps before they reach the primary Solver LLM [28]. This modularity is a key systems-level advantage; instead of retraining a trillion-parameter model, which is computationally prohibitive and ecologically unsustainable, we can optimize a smaller, more agile Refiner policy [17]. This inference-time paradigm allows for per-sample adaptation, where the reasoning trigger is tailored to the specific logical nuances of each individual query [28].

The training of these refiner policies involves a dual objective: solving the problem directly and refining candidate responses [13]. This "Generative Self-Refinement" skill is model-agnostic and generalizes to out-of-distribution tasks, making it a highly resilient component of the AI stack [22]. However, the system must be designed to avoid "reward hacking," where the Refiner learns to exploit the Solver's biases or leaks the answer into the prompt to artificially inflate scores [8]. To mitigate this, we propose the integration of an Adaptive Solver Hierarchy—a curriculum-based mechanism that aligns the difficulty of the environment with the Refiner's evolving competence, ensuring stable and honest learning trajectories [28].

4. Infrastructure and Deployment: Scalability and Robustness

Integrating self-correcting RL architectures into existing large-scale infrastructures requires careful consideration of computational trade-offs. Inference-time scaling—where the model

spends more "thinking time" to refine its output—improves performance on complex tasks but increases latency and resource consumption [3]. From a systems perspective, this necessitates a tiered deployment strategy. Low-stakes, high-volume queries can be handled by standard generative paths, while high-stakes reasoning tasks are routed through the self-correcting pipeline [28]. This "Route-LLM" approach optimizes the balance between cost and reliability.

Moreover, the sustainability of these systems is a growing concern. Training and maintaining massive models contributes significantly to carbon footprints and energy demand [15]. By focusing on modular RL architectures, we can achieve high-level reasoning performance without the need for constant, large-scale retraining [7]. This "Plan-then-Action" methodology allows for high-level planning guidance to be injected into the reasoning process, reducing the number of tokens generated and thus the overall energy per inference [7]. Such efficiencies are vital for the long-term viability of AI-integrated infrastructures.

Deployment robustness also hinges on the system's ability to handle adversarial inputs or unexpected environmental shifts. A self-correcting model that has been trained to detect its own inconsistencies is inherently more resilient to "jailbreaking" or prompt injection attacks that aim to derail its logic [12]. By internalizing the verification process, the model creates a defensive layer that evaluates the logical consistency of its own response against the user's original intent, effectively filtering out nonsensical or harmful reasoning paths before they are finalized.

5. Socio-Technical Implications and Governance

The deployment of reasoning LLMs is not merely a technical challenge; it is a socio-technical one that intersects with policy, ethics, and human trust [19]. As these models begin to function as "cognitive shortcuts" in social and professional settings, the definition of truth shifts from internal belief to architectural coupling and evaluation regimes [19]. Governance frameworks must adapt to this shift by mandating transparency in how reasoning chains are verified. If an AI system makes a recommendation in a healthcare or financial context, the underlying logic must be accessible and auditable by human regulators [21].

Fairness and bias also take on new dimensions in the context of self-correction. Reinforcement learning policies can inadvertently amplify biases present in the reward signals [8]. If the "verifiable rewards" are based on datasets that reflect historical inequalities, the self-correcting mechanism may refine the reasoning to align with those biases rather than objective logic [4]. Ensuring fairness requires a multi-agent perspective where diverse verification agents provide a pluralistic set of feedback signals [27]. This prevents the system from converging on a narrow, potentially biased "logical" path.

Policy-makers must also consider the implications of AI systems that can autonomously update their own "prompts and parameters" [4]. This level of agency necessitates new regulatory standards for "Agentic AI." We propose the development of "Reasoning Integrity Standards" that define acceptable error rates for logical consistency in critical sectors. These standards would move beyond simple accuracy metrics to evaluate the robustness of the reasoning chain itself,

ensuring that AI-integrated infrastructures remain stable and predictable even as the underlying models become more complex [27].

6. Forward-Looking Perspectives: The Path to Agentic Intelligence

Looking toward the future, the convergence of deep reinforcement learning and foundation models is reshaping the landscape of artificial intelligence [15]. We are moving away from fragmented, prompt-engineered agents toward unified architectural sciences [27]. In this new era, agentic intelligence is defined by the ability to manage and interact with "contextuality"—the dynamic integration of external observations and internal reasoning states [4]. Self-correcting RL architectures are the first step toward this more sophisticated form of agency.

We envision a future where LLMs operate within a "Contextual Cognition" framework, where reasoning is not a one-shot generation but a continuous process of perception, interaction, and refinement [4]. In this paradigm, logical hallucinations are treated as system faults that trigger automatic diagnostic and repair protocols. This evolution will likely involve "World-Model Pretraining," where models learn to reason by simulating outcomes in a structured environment before generating text [15]. Such systems would not just predict the next token; they would predict the consequences of their logic.

The ultimate goal is the creation of "Self-Evolving Agents" that can improve through continuous feedback without human intervention [4]. While this holds immense promise for scientific discovery and engineering innovation, it also requires a robust ethical and technical framework to ensure alignment with human values. The interdisciplinary researcher of the future must therefore be adept in both the high-dimensional mathematics of neural networks and the complex socio-technical dynamics of human-AI interaction.

7. Conclusion

The challenge of logical hallucinations in large language models represents a fundamental hurdle in the path toward reliable and trustworthy artificial intelligence. This paper has argued that the solution lies in the implementation of self-correcting reinforcement learning architectures that treat reasoning as a dynamic, modular process of refinement. By shifting the focus from static alignment to inference-time optimization, we can create systems that are more robust, scalable, and logically consistent.

However, the technical success of these architectures is inextricably linked to their socio-technical context. Robustness, fairness, and transparency must be designed into the system from the beginning, supported by a governance framework that understands the unique risks of agentic reasoning. As we integrate these "reasoning machines" into the fabric of our society, our responsibility is to ensure that their logic is not just a persuasive simulation, but a verifiable and grounded reflection of the world they are intended to serve. The refinement of reasoning chains is not just an engineering task; it is an essential step in securing the integrity of our future socio-technical infrastructures.

References

1. Alansari, A., & Luqman, H. (2025). Large Language Models Hallucination: A Comprehensive Survey. arXiv preprint arXiv:2510.06265.

2. Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5185–5198.

3. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

4. Cheng, M. (2026). A Comprehensive Survey of the LLM-Based Agent: The Contextual Cognition Perspective. Preprints.org.

5. Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., ... & Schulman, J. (2021). Training Verifiers to Solve Math Word Problems. arXiv preprint arXiv:2110.14168.

6. Deng, Y., Zhang, W., Chen, Z., & Gu, Q. (2023). Rephrase and Respond: Let Large Language Models Ask Better Questions for Themselves. arXiv preprint arXiv:2311.04205.

7. Dou, Z., Zhao, Q., Wan, Z., Zhang, D., Wang, W., Raiyan, T., ... & Biswas, S. (2025). Plan Then Action: High-Level Planning Guidance Reinforcement Learning for LLM Reasoning. arXiv preprint arXiv:2510.01833.

8. Gao, L., Schulman, J., & Hilton, J. (2023). Scaling Laws for Reward Model Overoptimization. International Conference on Machine Learning, 10949–10966.

9. Guo, Z., Han, Y., & Liu, X. (2025). Reinforcement Learning with Verifiable Rewards for Logical Consistency. Journal of AI Research, 78, 112–134.

10. Huang, M., Huang, R., Zheng, C., Li, J., Chen, G., Shi, H., & Cheng, H. (2025). Answer-Consistent Chain-of-thought Reinforcement Learning For Multi-modal Large Langauge Models. arXiv preprint arXiv:2510.10104.

11. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. ACM Computing Surveys, 55(12), 1–38.

12. Kim, S., Joo, H., Kim, J., & Lee, J. (2023). Recursive Introspection for Correcting Logical Fallacies in LLMs. NeurIPS Workshop on Socio-Technical AI.

13. Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., ... & Clark, P. (2023).

Self-Refine: Iterative Refinement with Self-Feedback. arXiv preprint arXiv:2303.17651.

14. Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in deterministic non-context-free grammar induction. Proceedings of the National Academy of Sciences, 117(48), 30046–30054.

15. Mienye, I. D. (2026). Deep Reinforcement Learning in the Era of Foundation Models: A Survey. Computers, 15(1), 40.

16. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

17. Paul, D., Ismayilzada, M., Peyrard, M., Beatson, I., & West, R. (2024). REFINER: Reasoning Feedback on Intermediate Representations for LLMs. arXiv preprint arXiv:2304.01904.

18. Piantadosi, S. T., & Hill, F. (2022). Meaning without reference in large language models. arXiv preprint arXiv:2208.02944.

19. Sariyar, M. (2026). Large language models as cognitive shortcuts: a systems-theoretic reframing beyond bullshit. Frontiers in Artificial Intelligence, 9, 1681525.

20. Shanahan, M. (2024). Role play with large language models. Nature, 623(7987), 493–498.

21. Tigard, D. W. (2025). The ethics of AI-generated bullshit. Ethics and Information Technology, 27(1), 14–22.

22. Wang, Q. (2025). SELF-REFINEMENT OF PARALLEL REASONING IN LLMS. OpenReview.

23. Wei, J., Wang, T., Schuurmans, D., Maarten, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Advances in Neural Information Processing Systems, 35, 24824–24837.

24. Xiong, W., Zhang, H., Ye, C., Chen, L., Jiang, N., & Zhang, T. (2025). Self-rewarding correction for mathematical reasoning. arXiv preprint arXiv:2502.19613.

25. Xu, B., Wang, L., & Zhang, Y. (2024). RE2: Reinforcement Learning for Reasoning Elicitation. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing.

26. Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., ... & Shi, S. (2025). Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. Computational Linguistics, 51(4), 1373–1412.

27. Zhou, C. (2026). From Fragmentation to Systematic Design: Architecting LLM-Based Multi-Agent Systems. TechRxiv.

28. Zhou, Y. (2026). Inference-Time Reasoning Elicitation via Reinforcement Query Refinement. arXiv preprint arXiv:2604.25444.